PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Monika Osińska
Uniwersytet Ekonomiczny w Poznaniu

MIERNIKI OCENY JAKOŚCI PODZIAŁU W ANALIZIE SKUPIEŃ – PORÓWNANIE ICH EFEKTYWNOŚCI

Streszczenie: Ocena jakości podziału jest jednym z czterech zasadniczych etapów analizy skupień, istotnie wpływającym na interpretację uzyskanych wyników. W literaturze przedmiotu istnieje duża liczba wskaźników wspomagających proces wyboru podziału optymalnego, jednak spora ich część przejawia pewne własności, które w dużej mierze mogą ograniczać obszary ich zastosowań. Głównym celem referatu jest zaproponowanie nowego wskaźnika oceny jakości grupowania – wskaźnika CNI oraz porównanie jego użyteczności z siedmioma najbardziej znanymi w literaturze wskaźnikami. Obiekty poddane analizie opisane zostały przy użyciu zmiennych dwuwymiarowych, natomiast kolejne podziały uzyskano w wyniku zaimplementowania metody *k*-średnich. Wszystkie obliczenia wykonano w programie R.

Słowa kluczowe: wskaźniki oceny jakości podziału, analiza skupień, klasyfikacja danych.

1. Wstęp

Procedura polegająca na ocenie wyników uzyskanych w drodze zastosowania wybranej metody klasyfikacji znana jest pod nazwą *oceny jakości podziału* (*cluster validity*). W literaturze przedmiotu istnieje bardzo duża liczba mniej lub bardziej efektywnych wskaźników pozwalających ustalić optymalną liczbę skupień. Spora ich część jednak ma pewne własności, które mocno ograniczają obszary ich zastosowań.

W niniejszym artykule dokonano analizy porównawczej wybranych wskaźników, zaproponowano także nowy wskaźnik oceny jakości podziału (CNI). Ze względu na możliwość prostej oceny wizualnej zastosowano zbiory obiektów opisane za pomocą zmiennych 2-wymiarowych, natomiast podziału tego zbioru dokonano za pomocą metody *k*-średnich, która jest obecnie najpowszechniej stosowaną metodą w tego typu analizach.

2. Wskaźniki oceny jakości podziału

W tabeli 1 zestawiono definicje wskaźników poddanych analizie porównawczej w dalszej części artykułu. Zakłada się przy tym, że m to liczba zmiennych opisujących obiekty, n to liczba obiektów, c − liczba skupień, n_i − liczebność *i*-tego skupienia, C_i − skupienie *i*-te (rozumiane jako zbiór obiektów), $x_k = \left(x_k^1, x_k^2, ..., x_k^m\right)$ − *k*-ty obiekt (rozumiany jako wektor), x_k^p − wartość *p*-tej zmiennej dla obiektu *k*-tego. Ponadto:

$$v_i = \left(\frac{1}{n_i}\sum_{k=1}^{n} u_{ki} x_k^1, \frac{1}{n_i}\sum_{k=1}^{n} u_{ki} x_k^2, ..., \frac{1}{n_i}\sum_{k=1}^{n} u_{ki} x_k^m \right),$$

$$v = \left(\frac{1}{n}\sum_{k=1}^{n} x_k^1, \frac{1}{n}\sum_{k=1}^{n} x_k^2, ..., \frac{1}{n}\sum_{k=1}^{n} x_k^m \right),\ u_{ki} = \begin{cases} 1 & \text{gdy } x_k \in C_i \\ 0 & \text{gdy } x_k \notin C_i \end{cases},$$

$$d\left(x_k, x_l\right) = \sqrt{\sum_{p=1}^{m}\left(x_k^p - x_l^p\right)^2},\ W\left(v_i\right) = \frac{1}{n_i}\sum_{k=1}^{n} u_{ki} d\left(x_k, v_i\right),$$

$$T\left(v\right) = \frac{1}{n}\sum_{k=1}^{n} d\left(x_k, v\right),\ W^2\left(v_i\right) = \frac{1}{n_i}\sum_{k=1}^{n} u_{ki} d^2\left(x_k, v_i\right),\ T^2\left(v\right) = \frac{1}{n}\sum_{k=1}^{n} d^2\left(x_k, v\right),$$

$$A(v) = \frac{1}{c}\sum_{i=1}^{c} d\left(v_i, v\right).$$

Tabela 1. Wskaźniki oceny jakości podziału – definicje

Nazwa indeksu	Definicja	Kryterium optymalizacji
1	2	3
Dunn's index [D]	$D = \min\limits_{i=1,..,c-1}\left(\min\limits_{j=i+1,...,c}\left(\dfrac{d\left(v_i; v_j\right)}{\max\limits_{i=1,...,c}\left(diam\left(C_i\right)\right)} \right) \right)$, $\forall\limits_{i=1,...,c} diam\left(C_i\right) = \max\limits_{x,y\in C_i}\left(d\left(x;y\right)\right)$	max
Modified Hubert Gamma Statistic [modH]	$\Gamma = \dfrac{2}{n\left(n-1\right)}\sum\limits_{k=1}^{n-1}\sum\limits_{l=k+1}^{n} d\left(x_k; x_l\right) q\left(x_k; x_l\right)$ $q\left(x_k; x_l\right) = d\left(v_i; v_j\right)$, gdy $x_k \in C_i$ oraz $x_l \in C_j$.	max
Root Mean Squared Standard Deviation index [RMSSTD]	$RMSSTD = \dfrac{\sum\limits_{i=1}^{c} n_i W^2\left(v_i\right)}{n-c}$	min.

Tabela 1, cd.

1	2	3
Root Squared index [RS]	$RS = \frac{\sum_{j=1}^{n} nT^2(v) - \sum_{i=1}^{c} n_i W^2(v_i)}{\sum_{j=1}^{n} nT^2(v)} = 1 - \frac{\sum_{i=1}^{c} n_i W^2(v_i)}{\sum_{j=1}^{n} nT^2(v)}$	max
Caliński Harabasz index [CH]	$CH = \frac{\frac{A}{c-1}}{\frac{W}{n-c}} = \frac{A(n-c)}{W(c-1)}$, $A = \sum_{i=1}^{c} n_i d^2(v_i, v)$, $W = \sum_{i=1}^{c} n_i W^2(v_i)$	max
Silhouette internal cluster quality index [S]	$S = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{k=1}^{n} u_{ki} S(k)$, $S(k) = \frac{b(k) - a(k)}{\max_{k=1,\ldots,n}\{b(k); a(k)\}}$, $a(k) = \frac{1}{n_i - 1}\sum_{\substack{l=1 \\ l \neq k}}^{n} u_{li} u_{ki} d(x_k; x_l)$, $b(k) = \min_{i \neq j}\{d(x_k; C_j)\}$ $d(x_k; C_j) = \frac{1}{n_j}\sum_{l=1}^{n} u_{ki} u_{lj} d(x_l; x_k)$, $i \neq j$	max
Davies & Boudlin's index [DB]	$DB = \frac{1}{c}\sum_{i=1}^{c}\max_{\substack{j \\ j \neq i}}\left(\frac{W(v_i) + W(v_j)}{d(v_i, v_j)}\right)$	min.

Źródło: opracowanie własne na podstawie: [Halkidi i in. 2002, 2001; Bezdek, Pal 1998; Gatnar, Walesiak 2004; Gusarova, Yatskiv 2005].

Zwykle procedura postępowania przy wnioskowaniu na podstawie tych wskaźników sprowadza się do analizy wzrokowej wykresów ich wartości. Niektóre z nich wykazują często tendencje do zmniejszania bądź zwiększania swoich wartości w miarę wzrostu liczby skupień, inne zaś tendencje do osiągania naprzemiennie raz wysokich, a raz niskich wartości, jeszcze inne często osiągają zbliżone wartości przy różnych podziałach. Stąd coraz częściej zaleca się korzystanie z innych dodatkowych kryteriów przy wyborze podziałów optymalnych [Najman 2007; Bezdek, Pal 1998].

3. Wskaźnik CNI

Ze względu na powyższe uwagi cel, jaki przyświecał konstrukcji opisanego dalej wskaźnika, to przede wszystkim jednoznaczność oraz trafność wskazań optymalnych podziałów.

Głównym zadaniem, a zarazem bardzo istotnym etapem konstrukcji wskaźnika CNI (Cluster Neighbourhood Index), jest wskazanie par skupień „sąsiadujących", tzn. takich, między którymi nie znajdują się elementy z żadnego z pozostałych skupień lub jest ich bardzo mało. Wartość skonstruowanego wskaźnika oblicza się zgodnie ze wzorem:

$$CNI = \frac{1}{r}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} o_{ij}\left(\frac{f^{\,i}\left(s_{ij}, p(i,j)\right)}{2n_i} + \frac{f^{\,j}\left(s_{ij}, p(i,j)\right)}{2n_j}\right),$$

gdzie: $p(i,j) = \max_{i,j}\left\{\bar{d}_i, \bar{d}_j\right\}$, $\bar{d}_i = \frac{1}{n_i}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} u_{ki}u_{li}d\left(x_k, x_l\right)$,

$$s_{ij} = \left(\frac{v_i^1 + v_j^1}{2}, \frac{v_i^2 + v_j^2}{2}, \ldots, \frac{v_i^m + v_j^m}{2}\right),$$

$f^{\,i}(s,p)$ – liczba obiektów ze skupienia *i*-tego, leżących wewnątrz sfery o środku w punkcie *s* i promieniu *p*,

r – liczba tzw. sąsiadujących skupień.

O tym, czy dane dwa skupienia są skupieniami sąsiadującymi, rozstrzyga funkcja o_{ij}:

$$o_{ij} = \begin{cases} 1 \text{ gdy } a_{ij} \le a \\ 0 \text{ gdy } a_{ij} > a \end{cases},$$

gdzie: $a_{ij} = \max_{\substack{l=1,\ldots,c \\ l\neq i \\ l\neq j}}\left(\max_{h=s_{ij}, c_{dij}, c_{gij}}\left(\frac{f^{\,l}\left(h, r(i,j)\right)}{n_l}\right)\right)$, $r(i,j) = \frac{\bar{d}_i + \bar{d}_j}{2}$,

$$c_{dij} = \left(\frac{s_{ij}^1 + v_i^1}{2}, \frac{s_{ij}^2 + v_i^2}{2}, \ldots, \frac{s_{ij}^m + v_i^m}{2}\right),\ c_{gij} = \left(\frac{s_{ij}^1 + v_j^1}{2}, \frac{s_{ij}^2 + v_j^2}{2}, \ldots, \frac{s_{ij}^m + v_j^m}{2}\right).$$

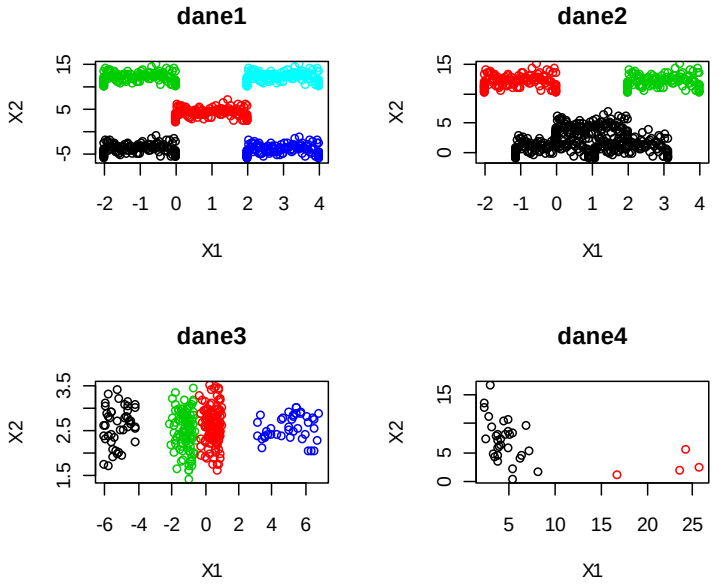
Problem, z jakim można się spotkać przy obliczaniu wartości tego wskaźnika, to ustalenie wartości *a*. Nie została ona w żaden sposób narzucana z góry. Jest to pewna przyjęta wartość graniczna, która pozwala wnioskować o sąsiedztwie skupień. W przeprowadzonych badaniach przyjęto $a = 0{,}2$, ponieważ dla tej wartości otrzymano najbardziej zadowalające wyniki.

4. Wyniki badań empirycznych

W niniejszej części artykułu przedstawiono wyniki przeprowadzonych analiz porównawczych. W celu pogrupowania danych zastosowano metodę *k*-średnich. Początkowe środki skupień zostały wyznaczone w wyniku zastosowania metody Warda. Autor zdaje sobie sprawę z tego, że efektywność metody w dużej mierze zależy od przyjętych kryteriów optymalizacyjnych [Najman 2008] oraz typów danych [Najman 2007], dlatego do analizy zostały wybrane tylko takie zbiory, dla których metoda trafnie zaklasyfikowała obiekty do wyodrębnionych skupisk.

Wybrano po cztery typy danych w ramach dwóch struktur: dobrze odseparowanych, w których odległości sąsiadujących obiektów wewnątrz skupienia są znacznie

mniejsze niż najmniejsza odległość obiektów z sąsiednich skupień, oraz słabo odseparowanych, w których odległości sąsiadujących obiektów wewnątrz skupienia są zbliżone do najmniejszej odległości obiektów z sąsiednich skupień (zob. rys. 1 i 2).

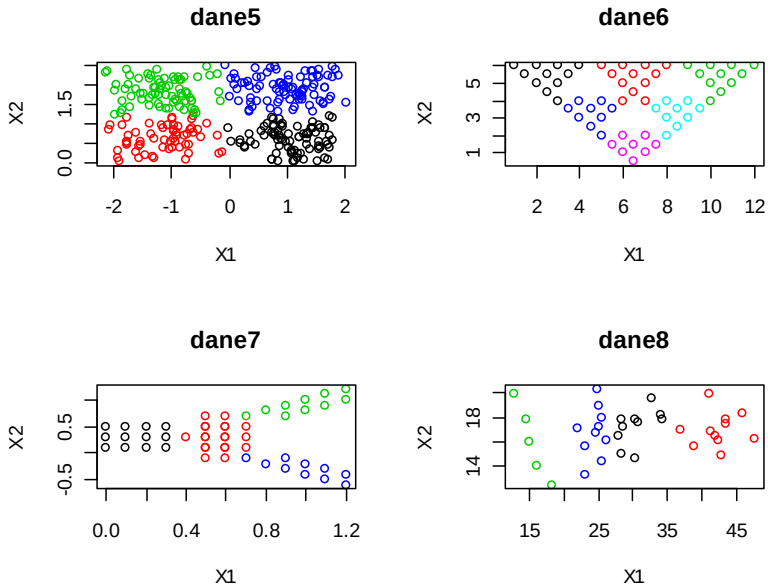


Rys. 1. Wykresy rozrzutu zastosowanych w analizie typów danych, rozpatrywanych w ramach struktur dobrze odseparowanych

Źródło: opracowanie własne przy wykorzystaniu programu R.

W tabeli 2 zamieszczono liczbę skupień, która została wskazana przez poszczególne indeksy. Wskazanie priorytetowe wyboru podziału optymalnego znajduje się poza nawiasem, tzn. przy danej liczbie skupień indeks osiągnął globalnie wartość optymalną w sensie zdefiniowanego kryterium, natomiast liczba skupień w nawiasach została wzięta pod uwagę ze względu na zbliżone do wartości optymalnej poziomy indeksu dla tych podziałów. W nawiasie przy wskaźniku CH podano optymalną liczbę skupień ze względu na kryterium największej lewostronnej zmiany wartości wskaźnika [Najman 2007].

Trafność wskazań została obliczona w następujący sposób: jeżeli dla danego indeksu wskazany przez niego podział priorytetowy jest (nie jest) podziałem optymalnym, to przypisuje się mu wartość odpowiednio 1 (0). Jeżeli natomiast podział optymalny jest jednym z podziałów alternatywnych, to przypisuje się mu wartość $1/l$, gdzie l jest liczbą wskazań podziałów optymalnych dla danego indeksu. Następnie liczy się średnią wartość dla poszczególnych typów danych w ramach każdego indeksu.



Rys. 2. Wykresy rozrzutu zastosowanych w analizie typów danych, rozpatrywanych w ramach struktur słabo odseparowanych.

Źródło: opracowanie własne przy wykorzystaniu programu R.

Wskaźniki najlepiej poradziły sobie z typem ‘dane3’ − wszystkie wskazały poprawnie podział optymalny. Są to dane separowalne liniowo w osi X o niejednolitej gęstości wewnątrz skupienia, dla których maksymalna odległość obiektów pochodzących z tego samego skupienia plasuje się na podobnym poziomie co odległości między centrami sąsiadujących skupień. Środkowe skupisko można dodatkowo rozbić na dwa słabo odseparowane skupienia (niektóre wskaźniki te skupienia wyodrębniły – zob. tab. 2).

Sporo trudności jednak przysporzył typ ‘dane2’, gdzie tylko jeden indeks (CH) poprawnie wskazał liczbę skupień, natomiast indeks CNI wskazał go jako jeden z podziałów alternatywnych. Dane te są separowane liniowo w osi Y o jednolitej gęstości wewnątrz skupienia. Maksymalna odległość obiektów pochodzących z tego samego skupienia jest jednak sporo większa od odległości między centrami sąsiadujących skupień.

Najwyższą trafność na poziomie 87,5% osiągnął proponowany indeks CNI, następnie miernik CH na poziomie 75% i Dunn – 62,5%. 56,3% trafności wykazały dwa wskaźniki: S oraz DB. Pozostałe wskaźniki (modH, RMSSTD i RS) miały trafność na poziomie 50%.

Wskazania poszczególnych indeksów dla typów danych rozpatrywanych w ramach struktur słabo odseparowanych zostały zestawione w tab. 3. Rozpatrywane indeksy najlepiej poradziły sobie z typem ‘dane6’, gdzie wszystkie (oprócz indeksu

Dunn) wskazały podział na sześć skupień jako podział optymalny priorytetowy, mimo że zbiory te wydają się dość problemowe ze względu na nieseparowalność tych danych względem żadnej z osi oraz jednostkowy stosunek maksymalnej odległości obiektów z tego samego skupienia do odległości między centrami sąsiadujących skupień.

Tabela 2. Zestawienie wskazań optymalnej liczby skupień przez poszczególne indeksy dla struktur dobrze odseparowanych

Wyszczególnienie	dane1	dane2	dane3	dane4	Trafność (%)
Liczba obiektów:	630	630	300	35	
Maksymalna liczba skupień:	10	10	10	6	
Rzeczywista liczba skupień:	5	3	3,4	2	
Dunn's index [Dunn]	3 (5)	2	3	2	62,5
modified Hubert Gamma [modH]	5	4	4	4	50,0
RMSSTD [RMSSTD]	5	4	4	4	50,0
RS [RS]	5	4	3	4	50,0
Caliński-Harabasz [CH]	5 (5)	3 (3)	4 (3)	4 (3)	75,0
Silhouette [S]	3 (2,4,5)	10 (9)	3	2	56,3
Davies-Bouldin [DB]	3 (2,4,5)	2	3	2	56,3
CNI [CNI]	5	2 (3)	3	2	87,5

Źródło: obliczenia własne przy wykorzystaniu programu R.

Tabela 3. Zestawienie wskazań optymalnej liczby skupień przez poszczególne indeksy dla struktur słabo odseparowanych

Wyszczególnienie	dane5	dane6	dane7	dane8	Trafność (%)
Liczba obiektów:	300	60	52	36	
Maksymalna liczba skupień:	10	8	8	6	
Rzeczywista liczba skupień:	4	6	4	3 lub 4	
Dunn's index [Dunn]	2	6,7,8	7,8	6	8,3
modified Hubert Gamma [modH]	?	6	5	4	50,0
RMSSTD [RMSSTD]	4	6	5	4	75,0
RS [RS]	4	6	5	4	75,0
Caliński-Harabasz [CH]	2 (4)	6 (6)	5 (3)	4 (3)	62,5
Silhouette [S]	2	6	3	2	25,0
Davies-Bouldin [DB]	2 (3)	6	3 (5)	2	25,0
CNI [CNI]	4	6	3	3,4	75,0

Źródło: obliczenia własne przy wykorzystaniu programu R.

Rozproszenie wewnątrz skupień w przypadku ‘dane7’ oraz poziom odseparowania i gęstość obiektów w skupiskach jest taka sama jak w przypadku ‘dane6’, ale trafność wskazań dla tego typu danych jest zerowa w ramach tych struktur. Najprawdopodobniej wynika to z elipsoidalnego kształtu skupisk − w przypadku ‘dane6’ skupiska są owalne.

Najwyższą trafnością wśród rozpatrywanych indeksów charakteryzują się wskaźniki CNI, RMSSTD oraz RS (75%). Na korzyść wskaźnika CNI, w porównaniu ze wskaźnikami RMSSTD oraz RS, przemawia jednoznaczność wskazań podziału optymalnego. Wskaźnik RMSSTD (RS) wykazuje tendencję odpowiednio malejącą (rosnącą) w zależności od liczby skupień. Wymusza to na badaczu przyjęcie innego (bardziej subiektywnego) kryterium ustalania podziałów optymalnych. W przypadku wskaźnika CNI obserwuje się nieznaczne pogorszenie trafności w porównaniu ze strukturami dobrze odseparowanymi, natomiast wskaźniki RMSSTD oraz RS są bardziej efektywne w odniesieniu do danych słabo odseparowanych. Stosunkowo wysoką trafność osiągnął również wskaźnik CH – 62,5%. Najgorsze okazały się wskaźniki: Dunn (8,3%), S (25%) oraz DB (25%).

Na koniec warto zwrócić uwagę na poziom trafności wskazań indeksu Dunn, który mocno osłabił swoją efektywność przy analizie struktur słabo odseparowanych. Potwierdza to siłę założenia, które zostało sformułowane w Migdał-Najman i Najman [2005] o tym, że odległości pomiędzy obiektami w skupieniu powinny być małe w stosunku do odległości między obiektami z różnych skupień.

5. Podsumowanie

Zaproponowany w niniejszym artykule wskaźnik oceny jakości podziału CNI pozwolił w sposób bardziej jednoznaczny, w porównaniu z pozostałymi siedmioma rozpatrywanymi indeksami, wskazać optymalną liczbę klas w analizowanych zbiorach obiektów. Charakteryzuje się on najwyższą trafnością, co oznacza, że najbardziej spośród tych wskaźników nadaje się do ustalenia optymalnej liczby skupień. Co więcej, niezależnie od tego, czy dane były dobrze odseparowane, czy słabo, wskazania indeksu CNI były jednoznaczne, a jego trafność była na porównywalnym poziomie.

Literatura

Bezdek J.C., Havens T.C., Keller J.M., Popescu M., *Dunn's Cluster Validity Index as a Constrast of VAT Images*, 19th International Conference on Pattern Recognition (ICPR),2008 December 8-11, Tampa, Florida, USA, IEEE.

Bezdek J.C., Pal N.R., *Some new indexes of cluster validity*, Transactions on Systems, Man, and Cybernetics, Part B", IEEE, 1998, vol. 28, no. 3, s. 301-315.

Gatnar E., Walesiak M., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wyd. Akademii Ekonomicznej, Wrocław 2004.

Gusarova L., Yatskiv I., *The methods of cluster analysis results validation*, Proceedings of International Conference RelStat '04, „Transport and Telecommunication" 2005, vol. 6, no. 1

Halkidi M., Batistakis Y., Vazirgiannis M., *On cluster validation techniques*, „Journal of Intelligent Information Systems" 2001, vol. 17, no. 2/3, s. 107-145.

Halkidi M., Batistakis Y., Vazirgiannis M., *Clustering validity checking methods: Part II*, „ACM SIGMOD Records" 2002, vol. 31, no. 3, s. 19-27.

Migdał-Najman K., Najman K., *Analityczne metody ustalania liczby skupień. Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 12, Akademia Ekonomiczna we Wrocławiu, Wrocław 2005, s. 265-273.

Migdał-Najman K., Najman K., *Analityczne metody ustalania liczby skupień w rozmytych zbiorach danych*. *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 13, Akademia Ekonomiczna we Wrocławiu, Wrocław 2006, s. 159-167.

Najman K., *Metody ustalania liczby skupień w zbiorach danych binarnych*. *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 14, Akademia Ekonomiczna we Wrocławiu, Wrocław 2007, s. 321-329.

Najman K., *Symulacyjna analiza wpływu wyboru kryterium optymalności podziału obiektów na jakość uzyskanej klasyfikacji w algorytmach k-średnich*. Klas*yfikacja i analiza danych – teoria i zastosowania*, Taksonomia 15, Uniwersytet Ekonomiczny we Wrocławiu, Wrocław 2008, s. 295-304.

THE CLUSTER VALIDITY INDICES – COMPARISON OF THEIR EFFICIENCY

Summary: Validation is one of four basic stages of cluster analysis which significantly affects the interpretation of the results. There are many validity indices in the specialist literature, but a lot of them have some properties which can strongly limit possible applications. The main goal of this paper is to propose a new validity index called *CNI* index and to compare it with the most popular indices. The data used in our analysis are described by 2-dimensional variables, while other divisions are achieved as a result of *k*-means algorithm implementation. All of the calculations were realized on the *R* programme.